A 128 × 128 SPAD Motion-Triggered Time of Flight Image Sensor with In-Pixel Histogram and Column-Parallel Vision Processor

A 128 × 128 SPAD Motion-Triggered Time of Flight Image Sensor with In-Pixel Histogram and Column-Parallel Vision Processor

Francesco Mattioli Della Rocca, *Student Member, IEEE*, Hanning Mai, *Member, IEEE*, Sam W. Hutchings, Tarek Al Abbas, *Member, IEEE*, Kasper Buckbee, *Student Member, IEEE*, Andreas Tsiamis, Peter Lomax, Istvan Gyongy, Neale A. W. Dutton, *Member, IEEE* and Robert K. Henderson, *Senior Member, IEEE*

***Abstract*—A 128 x 128 SPAD motion detection-triggered time of flight (ToF) sensor is implemented in STMicroelectronics 40 nm CMOS SPAD foundry process. The sensor combines vision and ToF ranging functions to acquire depth frames only when inter-frame intensity changes are detected. The 40 µm x 20 µm pixel integrates two 16-bit time-gated counters to acquire ToF histograms and repurposes them to compare two vision frames without requirement for additional out-of-pixel frame memory resources. An embedded column-parallel ToF and vision processor performs on-chip vision frame comparison and binary frame output compression as well as controlling the time-resolved histogram sampling. The sensor achieves a maximum 32.5 kfps in vision modality and 500 fps in motion detection-triggered ToF over a measured 3.5 m distance with 1.5 cm accuracy. The vision function reduces the sensor power consumption by 70% over continuous ToF operation and allows the sensor to gate the ToF laser emitter to reduce the system power when no motion activity is observed.**

***Index Terms*—3D imaging, CMOS, time of flight (ToF), histogramming, image sensor, light detection and ranging (LIDAR), single-photon avalanche diodes (SPADs), vision, motion detection**

I. Introduction

THE Internet of Things (IoT) promises networks of sensors seamlessly exchanging information for the sensing and actuation of connected devices with low power consumption and event-driven data rates. State of the art vision cameras have overcome these challenges providing high-frame rate motion detection imaging at low power consumptions [1]-[7] by selectively reading out and processing frames or pixels when intensity changes are detected. Multi-modal motion detection vision cameras have also been proposed to rapidly switch from a low power and reduced data readout regime for idle activity to high resolution intensity imaging after detection of a motion event [8]-[10]. These vision techniques however have only been applied to 2D imaging, and vision sensors do not currently provide depth information. On the other hand, light detection and ranging (LiDAR) cameras for 3D imaging operate continuously, consuming high power on the sensor and emitter and transferring at high data rates regardless of motion activity in the scene. 3D imaging applications such as object detection in robotic automation, traffic monitoring, person tracking in surveillance for security and assisted living, all feature scenarios where the sensor idle observation time can largely exceed the event time and event frequency. By limiting the 3D imaging operation duty-cycle to short time intervals triggered by the occurrence of an event, LiDAR image sensors targeting event-based applications will be able to deliver 3D information efficiently in an IoT network.

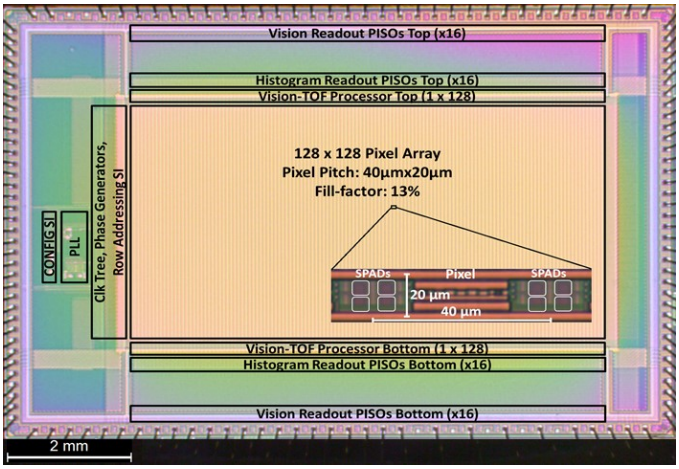

Fig. 1: Micrograph of the image sensor IC with pixel inset.

Motion detection functionality was first integrated in a 3D camera in [11]. Ranging was performed by combining the techniques of structured light stripe scanning and triangulation. The method however required a high frame rate and complex computation to obtain the range measurement, resulting in a high sensor power consumption in the order of Watts. The

Manuscript received November 30, 2019. This work was supported in part by STMicroelectronics through the funding of the PhD of F. Mattioli Della Rocca and in part by the Biotechnology and Biological Sciences Research Council (BBSRC) under Grant BB/R004226/1. *(Corresponding author: Francesco Mattioli Della Rocca)*

F. Mattioli Della Rocca, H. Mai, S. W. Hutchings, K. Buckbee, A. Tsiamis, P. Lomax, I. Gyongy and R. K. Henderson are with the School of Engineering, The University of Edinburgh, Edinburgh EH9 3JL, U.K. (e-mail: francesco.mattiolidellarocca@ed.ac.uk).

T. Al Abbas was with the School of Engineering, The University of Edinburgh, Edinburgh EH9 3JL, U.K. He is now with Sense Photonics, Edinburgh, U.K.

N. A. W. Dutton is with STMicroelectronics, Edinburgh EH3 5DA, U.K.

presence of scanning hardware also made the method complicated to integrate into a low-power 3D imaging system.

Flash LiDAR based on the principle of time of flight (ToF) is the least computationally demanding technique amongst other ranging methods such as triangulation, interferometry and structured light, and the flood illumination eliminates the mechanical complexity of a scanning emitter. Despite these system advantages, the integration of ToF image sensors in IoT nodes or mobile devices is challenging due to their data volumes exceeding hundreds of kilobytes per frame [12]-[15] to deliver sub-centimeter distance precision at high frame rate and spatial resolution. ToF cameras have been proposed that histogram time-correlated single photon counting data at each pixel to pre-process photon time stamps [12] and in doing so reduce the output data volume for each frame. Other SPAD ToF cameras can output direct depth maps [16] or embed histogram processing to output the position of the peak [13]. Despite these compression techniques, ToF systems still read out high spatial and temporal resolution data for every frame irrespective of scene activity. This continuous operation in ToF systems results in high power consumption due to the uninterrupted triggering of the laser emitter [14] [15], generation of high frequency time-gates in-pixel and the continuous transfer of ranging data to an external processor, the highest power contributors in a ToF system [15] [17].

In this paper, we propose a scheme where the ToF camera is triggered upon motion detection allowing a reduction in the system power by gating the laser emitter and by avoiding readout and processing of high resolution ToF frames with no motion activity. An embedded column-parallel processor performs vision frame comparison on-chip, reading out binary frames signaling the presence or absence of inter-frame activity. Once a vision event is registered, the camera is seamlessly switched to ToF operation and captures a time-resolved histogram from in-pixel photon sampling time gates [18]. In this way, only ToF frames containing motion information are ever read out for processing, thus reducing the data transferred during idle operation and eliminating the power consumption contribution of the ToF laser emitter and external processor in the absence of motion activity.

II. Sensor Architecture

A block diagram of the sensor overlaid on the chip micrograph is shown in Fig. 1. The imager is fabricated in STMicroelectronics 40 nm CMOS foundry technology optimized for SPADs [19] and comprises a 128 x 128 array of pixels. Each half-column of the imager array is mapped to a corresponding Vision-ToF processor as shown in Fig. 2 with top and bottom readout. The array can be operated in rolling or global shutter exposure. The column-parallel processors sample pixel frames on a rolling row-by-row scheme. A PLL generates a global high-frequency clock with range 500 MHz-1 GHz. This clock or a divided version of it (÷2/4/8/16) is vertically distributed to the left side of the array via a clock tree. The high-frequency clock is then further divided into 12 edge-shifted clock phases by phase clock generator blocks composed of row-parallel shift registers, one block shared by two rows, as shown in Fig. 3. The 12 phases are then distributed horizontally across the imaging array at alternate rows, two rows of pixels sharing the same clock phase lines. The clock phases are used by the pixels to generate dual time-gates for time-resolved in-pixel histogram sampling of SPAD events. A configuration shift register (CONFIG SI) controls the operating mode of the sensor. A total of 32 parallel input to serial output (PISO) shift registers read out either the vision or time of flight data, 8 imager columns sharing one readout serial output data line.

A. Pixel Architecture

A diagram of the pixel is shown in Fig. 4. The pixel is 40 µm by 20 µm in pitch with 13% fill factor similar to the one presented in [20]. Each pixel comprises 4 passively quenched SPADs split either side of the pixel electronics. The outputs of the SPAD circuits can be selectively masked to avoid sampling

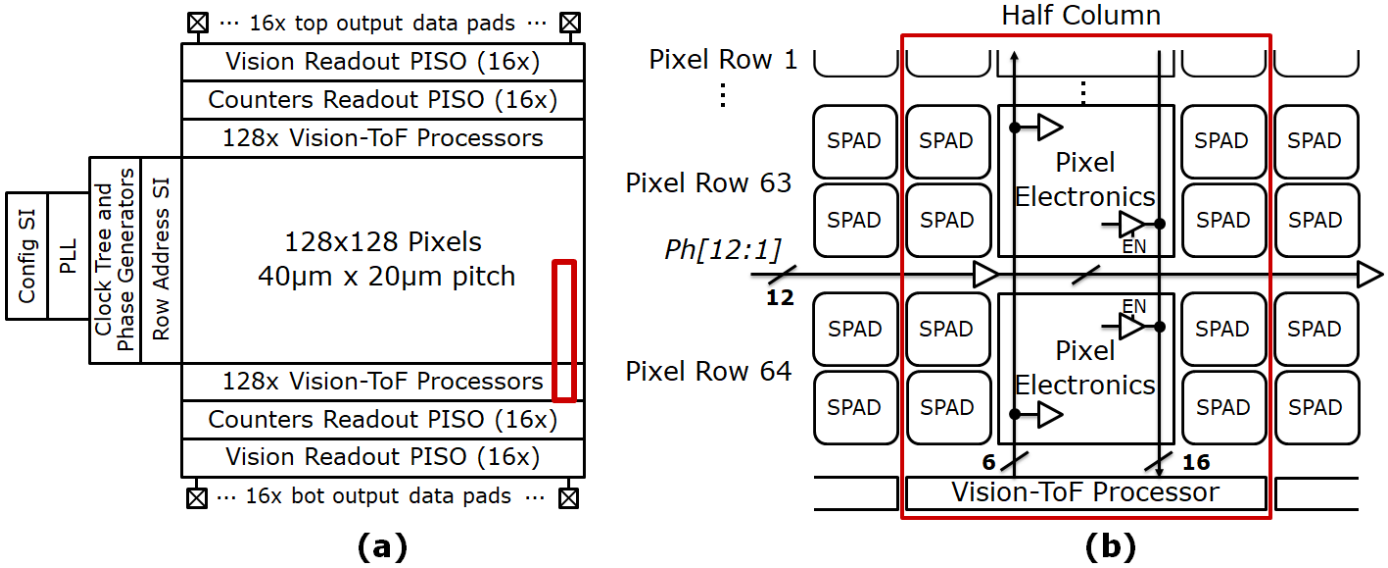

Fig. 2: (a) Block diagram of the sensor architecture with half-column inset. (b) Half-column of pixels interfacing to column-parallel Vision-ToF processor.

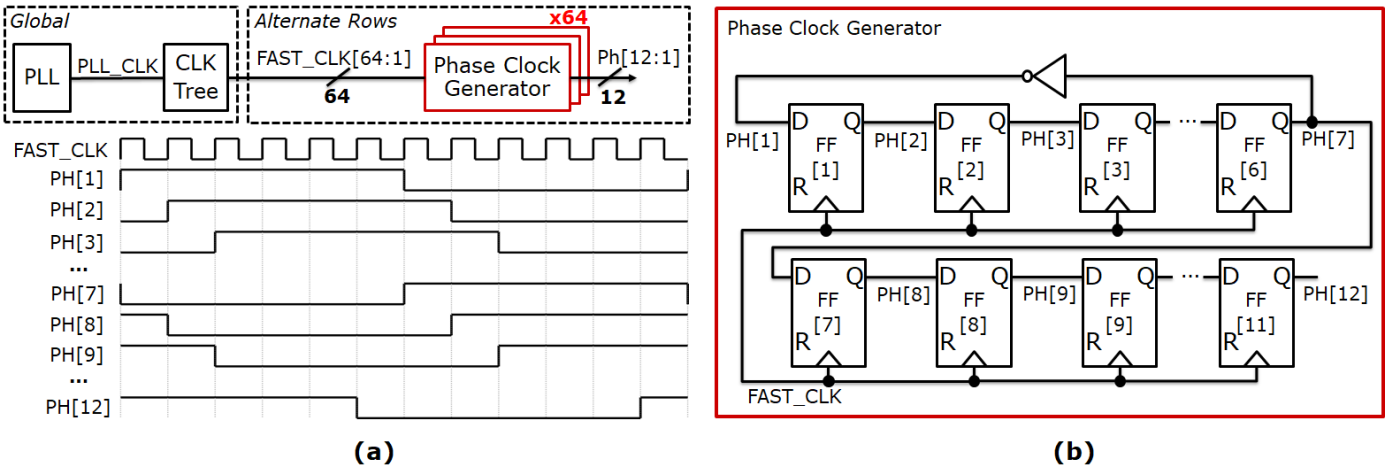

Fig. 3: (a) Block diagram and timing diagram of the generation of the phase clocks. (b) Schematic of the phase clock generator at the edge of alternate rows.

events from high dark count SPADs by writing to a 4-bit in-pixel memory. Pairs of SPADs from neighboring pixels are arranged in a column-wise well-sharing layout. The pitch of the pixel electronics matches the SPAD pitch for compatibility to stacked processes. The digital output pulses from the SPADs are shortened and combined by an OR tree into a single stream of event pulses. From the 12 clock phases arriving to the pixel, 3 clocks are selected to generate two time-gates, TG1 and TG2, each spanning two clock phase shifts. The programmable width of the time-gates can therefore range between 2 ns-64 ns according to the PLL frequency and clock division setting. The 2 gates can be scanned across the temporal range in steps of 1 phase shift by writing to a 6-bit in-pixel phase selection memory storing the selection of the clock phase centered between the two time gates. SPAD events are quantized into either time-gate and two 15-bit counters count the events occurring within the respective time-gate. An additional bit for each counter locks the SPAD sampling to avoid counter rollover.

While most frame-comparison vision cameras require an out-of-pixel frame memory to store the previous frame, in this sensor the time-gated counters for ToF operation are repurposed in the vision modality to store consecutive intensity frames for processing by the embedded column-parallel processors.

B. *Vision-Driven ToF*

The Vision-ToF processor is a digitally synthesized and automatically place-and-routed logic block occupying an area of 40µm by 80µm matching the horizontal pitch of one pixel column. The logic integrated in the processor is shown in Fig. 5. The processor enables the camera to operate in three different modalities: ToF for 3D imaging, vision for motion detection, and vision-driven ToF for motion-triggered 3D imaging, hereafter known as MD-ToF mode. The following section will outline the operation of these three modalities.

In vision mode the pixel toggles sampling of successive frame exposures in either one of the two counters. A diagram of the vision processing operation is shown in Fig. 6. A timing diagram detailing the vision operation is shown in Fig. 7. A first frame *i* is acquired by all pixels using the same counter, for

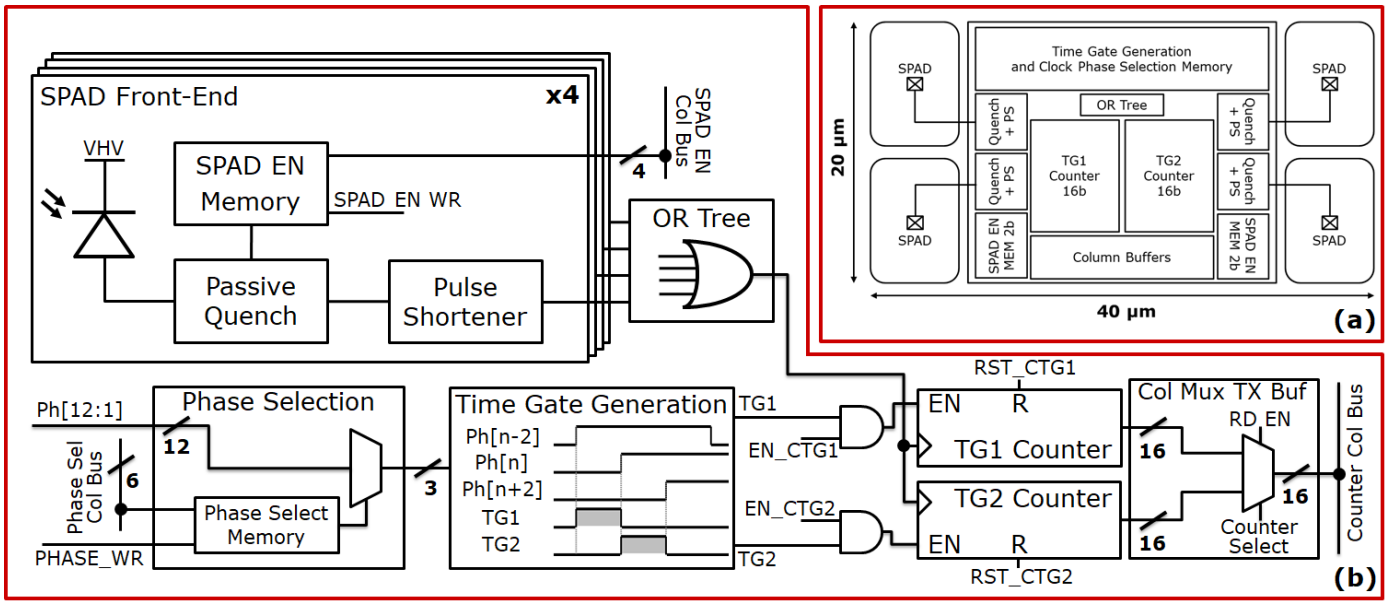

Fig. 4: (a) Pixel top-level block diagram. (b) Block diagram of in-pixel circuits.

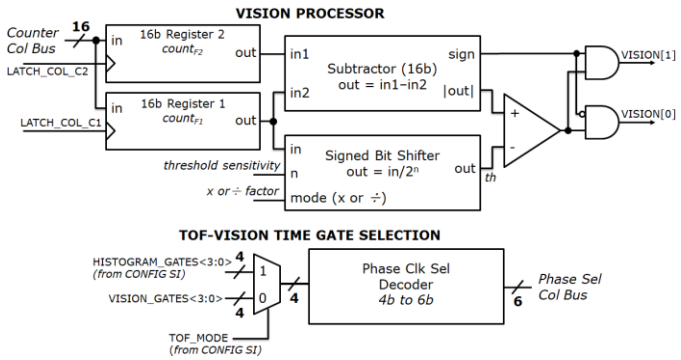

Fig. 5: Digital logic blocks integrated in column-parallel Vision-ToF processor.

example TG1 counter. The counter is reset at the beginning of the exposure by raising signal RST_CTG1 followed by exposure signal EN_CTG1 gating the counter on. During exposure *i,* TG2 counter stores the photon counts from the previous exposure *i-1* so its exposure gate EN_CTG2 remains low gating the counter off. At the end of exposure *i*, rows of pixels sequentially read the counter values into the column-parallel Vision-ToF processors, which compare the counts from frame *i* to frame *i-1*. At the start of the next exposure *i+1*, TG1 counter is not reset and retains the photon counts from frame *i* during acquisition of frame *i+1*. TG2 counter is reset by triggering RST_CTG2 and is used by the pixels to capture frame *i+1* by raising exposure signal EN_CTG2. After exposure *i+1* the processor again compares the counts between frame *i+1* and frame *i.* TG1 counter is then reset and captures frames *i+2* while TG2 Counter retains frame *i*+1. The pixel counters continue alternating between acquisition of a new frame and storage of the previous frame so that frame comparison can be performed at the end of every exposure by the processor. No additional frame buffer memory other than the pixel time-gated counters is required for storage of the previous frame.

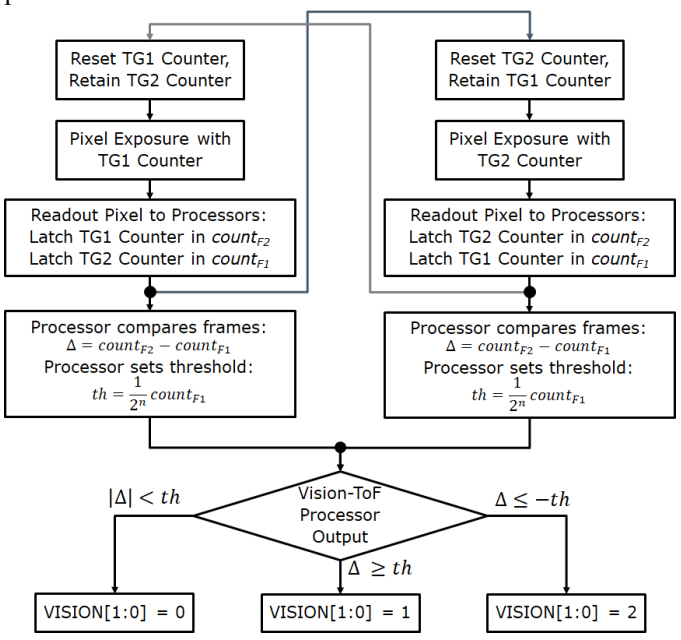

Fig. 6: Flow diagram of the sensor in vision frame comparison operation.

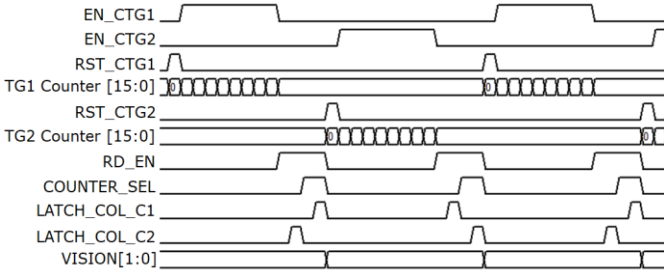

Fig. 7: Timing diagram of the vision modality.

The two counter values are read into the processor sequentially by row at the end of every exposure into two 16-bit registers. A two's complement signed subtractor integrated in the processor calculates the difference between the two consecutive frames Δ by subtracting the count of the former frame ($count_{F1}$) from the latter frame ($count_{F2}$) as in Eq. 1.

$$\Delta = count_{F2} - count_{F1} \tag{1}$$

The photon count from the former of the two frames is used to set the intensity change threshold (*th*). The processor calculates the threshold at each pixel for a valid vision event to be a factor of $count_{F1}$ as shown in Eq. 2:

$$th = \frac{1}{2^n} count_{F1} \tag{2}$$

where *n* is a globally configurable integer coefficient called the threshold contrast sensitivity coefficient and controls the relative change in brightness required to sense a valid vision event. It is programmed by a 4-bit register in CONFIG SI with values ranging from -7 to 7. A bit-shifter in the processor implements the multiplication and division by a power of two.

The vision threshold at each pixel therefore changes dynamically and is proportional to the photon activity at each pixel [1]. The difference between the two counters Δ is compared against the threshold, and the processor outputs one of three possible VISION output logic states: 0 to indicate that $-th < \Delta < th$, corresponding to no significant brightness change, 1 for the case that $\Delta \geq th$, corresponding to the current frame detecting an increased intensity from the previous frame, and 2 if $\Delta \leq -th$, corresponding to a decreased intensity. Expanding the above inequalities using Eq. 1 and Eq. 2 yields Eq. 3-5 for the three VISION processor output conditions:

$$\text{VISION} = 0 \text{ if } \quad -\frac{1}{2^n} < c < \frac{1}{2^n} \tag{3}$$

$$\text{VISION} = 1 \text{ if } \quad c \geq \frac{1}{2^n} \tag{4}$$

$$\text{VISION} = 2 \text{ if } \quad c \leq -\frac{1}{2^n} \tag{5}$$

where $c = \frac{count_{F2} - count_{F1}}{count_{F1}}$ is the temporal contrast and measures the relative change in light intensity between successive frames necessary for a vision event at a pixel to be detected as motion in the scene. The coefficient *n* therefore controls the contrast sensitivity to motion events according to Eq. 3-5.

State of the art event-based vision sensors conventionally use a logarithmic pixel to be sensitive to relative changes in light intensity rather than absolute changes [1]. This improves their resistance to photon shot noise, reducing false positives, and allows detection of brightness changes over a high dynamic range. The digital counters within the pixel of this sensor do not allow acquisition of light on a logarithmic scale. To circumvent this limitation, logarithmic sensitivity is instead implemented by the Vision-ToF processors at column-level by computation of the adaptive per-pixel threshold *th.* Expanding the inequalities in Eq. 3-5 and applying the logarithm on both sides yields Eq. 6-8.

$$\text{VISION} = 0 \text{ if } j < \log(count_{F2}) - \log(count_{F1}) < k \quad (6)$$
$$\text{VISION} = 1 \text{ if } \log(count_{F2}) - \log(count_{F1}) \geq k \quad (7)$$
$$\text{VISION} = 2 \text{ if } \log(count_{F2}) - \log(count_{F1}) \leq j \quad (8)$$

where $j = \log\left(-\frac{1}{2^n} + 1\right)$ and $k = \log\left(\frac{1}{2^n} + 1\right)$ for $n > 1$. If $n \leq 0$, only vision outputs 0 and 1 are possible, as reductions in relative brightness by more than a factor of one are not valid. This processing method allows the logarithmic vision algorithm to be implemented in the digital domain yielding the motion detection properties similar to an analogue logarithmic pixel [2] [3] [4].

The vision output of each pixel is encoded in a 2-bit number. An additional bit per pixel used for other processor functions and set to 0 in vision mode is serially read out with the 2-bit vision output for a total of 3 bits per pixel. The data is serially read out through the 32 output serial lines at 50 MHz clock rate corresponding to a maximum frame rate of 32.5 kfps. In practice, for motion detection the sensor is operated at much lower frame rates (10-1000 fps), to be able to accumulate enough photons for noiseless frame comparison, as well as reducing power consumption resulting from the low data rate at lower frame rates.

In the ToF modality the sensor performs a sequential scan of the pixel bins reading out 3 frames for a total of 6 histogram bin photon counts. A timing diagram of the operation of the ToF modality is shown in Fig. 8. For each frame, pixels are configured by the Vision-ToF processor to globally shift TG1 and TG2 in every pixel by two clock phase edges, thus covering the full temporal range of the histogram in 3 frame readouts. Alternatively, the sensor can be configured for dual-bin indirect ToF (IToF) ranging by generating TG1 and TG2 to cover the entire temporal range and reading out the 2 ToF bins in a single frame.

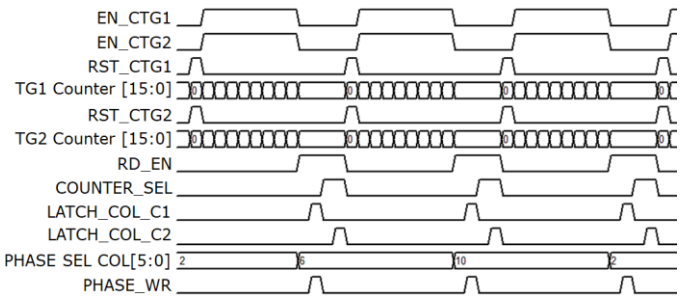

Fig. 8: Timing diagram of the ToF modality.

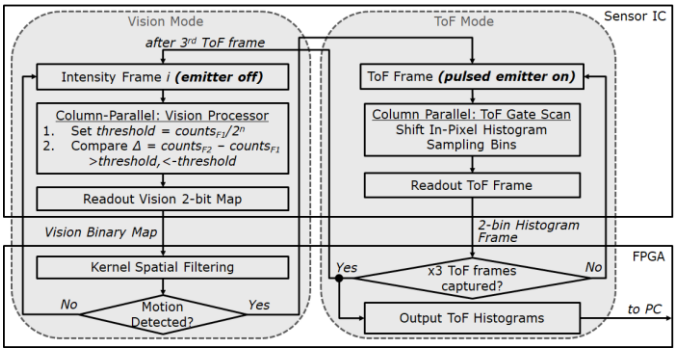

Fig. 9: Flow diagram of the MD-ToF modality.

In both vision and ToF modes, vision bit-map frames and histogram bin counts respectively are read out to an FPGA and passed on to a PC for processing of the vision and depth-map images.

The MD-ToF mode is used to operate the sensor in an always-on lower power vision mode until a motion event is detected. A flow diagram of the operation of this mode is shown in Fig. 9. The sensor starts its operation as in the vision modality, reading out motion bit-maps to the FPGA. Spatial filtering of the bit-map is performed on the FPGA with a programmable kernel size, checking the output of neighboring pixels to filter spurious false positives. If all pixels in the same kernel output a state of 1 or 2, the FPGA immediately triggers the sensor to acquire a ToF 6-bin histogram by acquiring three consecutive histogram frames as in the ToF modality. After the ToF frame has been acquired, the sensor is triggered to revert to scanning the scene in vision mode until the next vision event.

The sensor controls the laser trigger for ToF operation. The laser trigger is masked during vision operation, thus saving power on the pulsed laser emitter. In vision mode, the PLL clock is switched to its lowest frequency and divider setting, outputting a 30 MHz clock to reduce the power consumed in the distribution of the high frequency phase clocks for in-pixel gate generation, only necessary for ToF operation. Exposures can be independently set for vision and histogram acquisition, thus vision can be operated at a slower frame rate for further power saving during MD-TOF motion detection.

III. Sensor Characterization

This section will present the results of the characterization of the three sensor imaging modalities. The ultimate goal of the sensor is to measure scene depth only in presence of motion. The first part of this section describes the performance of the ToF-based 3D imaging function of the sensor. The second part of the section characterizes the motion detection operation of the sensor, presenting the results obtained by testing the frame comparison-based vision sensor function. The third part of the section will test the motion-triggered ranging sensor modality, which combines the previous ToF and vision modes to capture depth maps only when triggered by motion in the scene.

A. 3D Imaging

Mismatches in the in-pixel time gates would result in non-linearity in the range measurements. The profile of the time

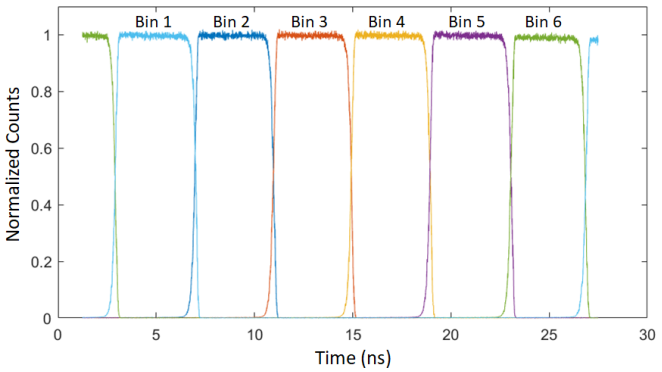

Fig. 10: Timing profile of the in-pixel time gates for ToF histogram bin sampling. Counts normalized to average photon count rate.

gates was therefore measured. The profiles of the ToF time-gates of the sensor were characterized by delaying a Hamamatsu picosecond laser (PLP-10, 483 nm, 70 ps FWHM) in 5 ps time steps across the histogram timing range. Histogram frames were captured of the laser pulse for each delay step. As the laser pulse enters and leaves each histogram bin it outlines the rising and falling edges of the 6 time gates. The DNL is calculated based on mismatches in the measured widths of the time gates. The timing profile of the time-gates is shown in Fig. 10. The six gates were generated with 4 ns FWHM (PLL operating at 500 MHz output clock) with a worst case DNL averaged across the pixel array of 80 ps equivalent to 2% of the time gate width.

The clock phases are horizontally distributed across the array from the left to the right hand side of the array. Due to the propagation delay of the clock phases this will result in a depth offset from one column to the next. The time delay of the time-gates horizontally across the pixel array is shown in Fig. 11. The delay was measured by shining a collimated laser beam (Picoquant FSL500, LDH-S-C-840) with 6 ns FWHM pulse width and 840 nm wavelength at the sensor and calculating the centre of mass of the pulse from the resulting histograms acquired in the ToF mode of the sensor. The delay offset between the first and last column is 6.58 ns with a 10.9 ps standard deviation across rows. This measurement is used to post-process sensor images, correcting for depth offset across the array.

The ranging performance of the sensor was evaluated by measuring the distance of a physical target using a Picoquant pulsed laser (840 nm, 6 ns FWHM) under dark and ambient lighting conditions with 500 lux background illumination. The sensor was again operated in ToF mode with 4 ns time gates

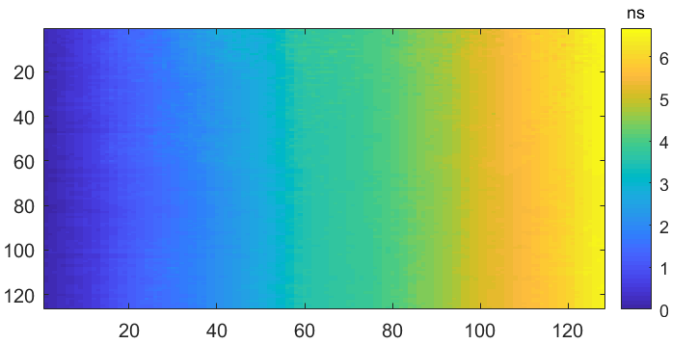

Fig. 11: Time delay of the time-gates horizontally across the pixel array.

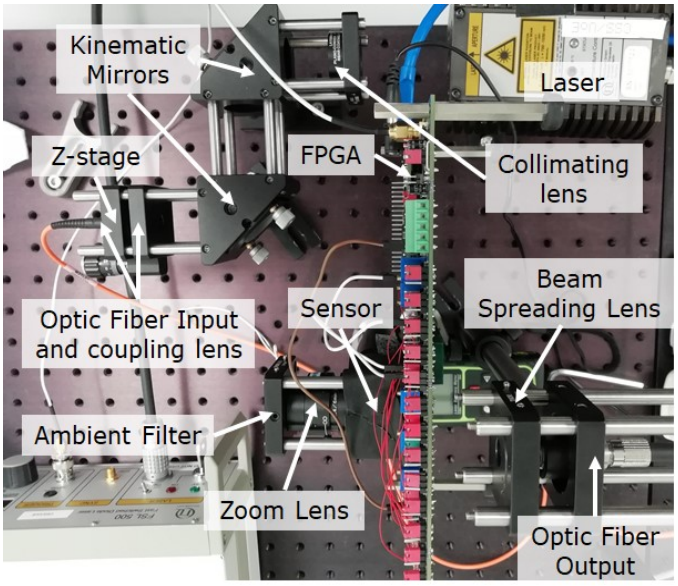

Fig. 12: Optical experimental setup for flash ToF illumination with fiber-coupled laser and focusing of reflected light from scene onto sensor.

(PLL operating at 500 MHz output clock). The target was placed from 0.5 m to 3.5 m away from the sensor, stepping it in increments of 25 cm. The Picoquant pulsed laser was fiber-coupled and the beam at the output of the fiber, measuring an optical power of 7 mWcm^{-2}, was spread with a 40˚ field of view (FOV) towards the target. The reflected light was focused onto the sensor focal plane with a zoom lens with matching FOV. The optical experimental setup is shown in Fig. 12. The distance was measured with 10 ms exposures. The measured distance, precision and non-linearity errors are shown in Fig. 13. The precision represents the standard deviation of the distance over 100 measurements and the non-linearity is the difference between the range measurements and the actual distance (ground truth). The sensor achieves a 1.5 cm rms non-linearity error over a 3 m range and a precision ranging from 1.8 cm at the shortest distance to 6.4 cm at the longest distance.

To demonstrate 3D imaging with the sensor, a depth map of a scene was captured and is shown in Fig. 14. The 3D image shows measurements of scene depth features down to centimeter accuracy. Due to the 40 µm × 20 µm pixel

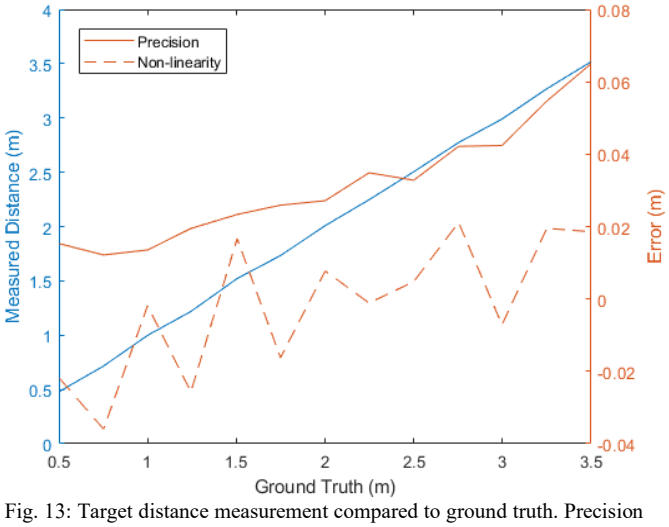

Fig. 13: Target distance measurement compared to ground truth. Precision and non-linearity errors are plotted on right vertical axis.

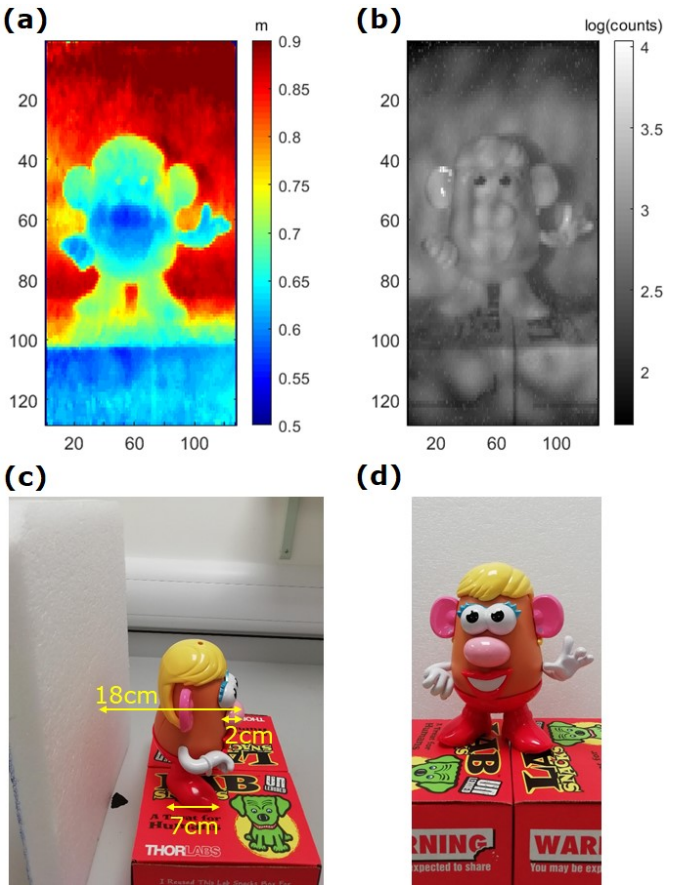

Fig. 14: (a) Depth map of toy captured at 40 fps. (b) Intensity image captured by sensor. (c) Reference image from side showing depth of scene. (d) Reference intensity image from standard camera.

dimensions, output images are stretched by a 2x1 aspect ratio. The images are therefore resized before being displayed to account for the pixel dimensions.

The same flash LiDAR optical setup is used as in the target range measurement. As shown in the setup in Fig. 12, an ambient filter is used to suppress background light in ToF acquisitions. The vision modality is however based on acquisition of passive illumination, so blocking the background completely would adversely impact the vision sensitivity at low intensities. A solution to this challenge is a trade-off between background rejection in ToF and background capture in vision by using a band-pass filter centered around the laser wavelength with a wider pass-band than conventional narrow band-pass filters. The ambient filter used throughout our ToF and vision experiments is an 800 nm filter with pass-band from 600 nm to 1000 nm allowing a high signal to background ratio for 3D imaging, blocking most of the high intensity visible range, while still passing enough light in both indoor and outdoor lighting conditions to maintain vision sensitivity at low light levels. An alternative solution could use wafer-level filters [21] on certain SPADs within the 2x2 macro-pixel. As individual SPADs can be masked, half of the SPADs in the pixel with IR filters could be used for ToF 3D imaging acquisition while the other SPADs with ambient light sensing (ALS) filters could be activated during vision motion detection.

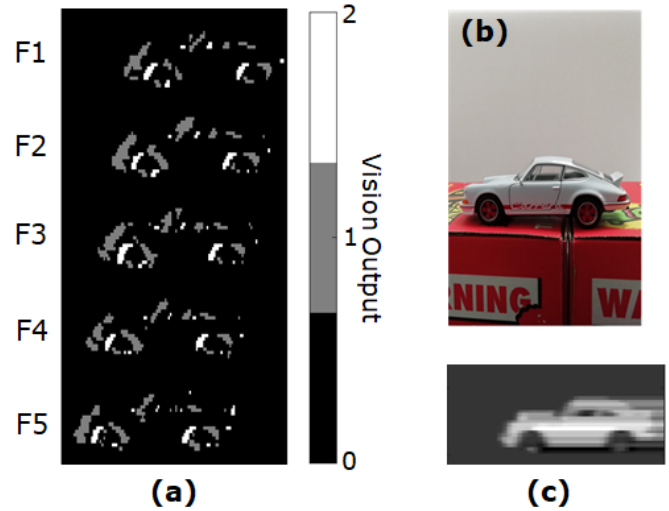

Fig. 15: (a) Five vision bit-map frames from 100fps video of a travelling toy car captured in the sensor vision mode. (b) Reference intensity image from standard camera showing captured toy car over white background. (c) intensity reconstruction based on motion vector estimation from 10 vision frames, and cumulative sum of averaged (realigned) vision frames along the direction of motion [22].

B. Motion Imaging

The motion detection imaging function of the sensor was tested by capturing a video of a moving toy car travelling across the scene. The video was captured at 100 fps operating the sensor in vision mode. Vision bit-map frames from the video showing the vision output of the sensor in response to the motion of the car are shown in Fig. 15. The white car is imaged over a white background providing a noise-challenging scenario. The vision frames show car edges displaying vision outputs 1 and 2 in line with the changes in intensity as the car drifts across the scene. Vision motion outputs within the car are also captured between the side windows, wheels and the car body corresponding to changes in color at different sections of the toy car. Fig. 15c shows an intensity reconstruction based on motion vector estimation from 10 vision frames, and cumulative sum of averaged realigned vision frames along the direction of motion [22].

A second video was captured of a spinning disk made of gray-scale sections, with 4% incremental percentage darkness over white, interspaced by white sections as shown in Fig. 16. The experiment allows verification of vision sensitivity over different contrast intensity changes. Vision binary frames show the sensor detecting changes at each section transition for

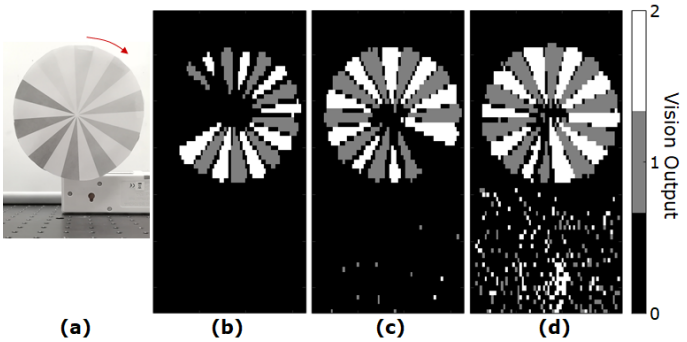

Fig. 16: Vision binary frames of rotating disk with 4% incremental contrast sections (white to gray) captured at different sensor contrast threshold sensitivities: (a) reference intensity image taken from standard camera (b) sensitivity to high contrast ≥12.5% (n=3), (c) sensitivity to medium contrast ≥6.25% (n=4), (d) sensitivity to low contrast ≥3.1% (n=5).

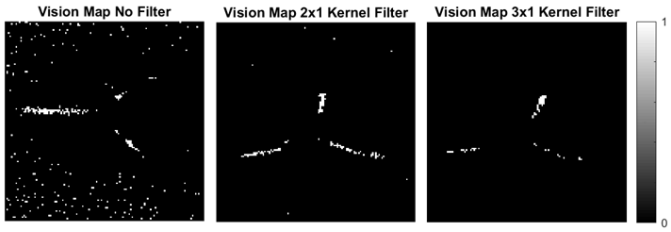

Fig. 17: Vision bit-maps from a rotating 3-blade fan with no spatial filtering (left), 2x1 kernel spatial filtering (middle), and 3x1 kernel spatial filtering (right).

different values of threshold sensitivity to contrast *n*. The spatial smearing of motion detection events across pixels observed at the section transitions is due to the difference between the rotating disk speed of 1200 rpm and the sensor frame rate used in the experiment of 500 fps. Higher values of *n* result in the camera increasing sensitivity to lower contrast transitions (according to Eq. 3-5). For example, setting n=3 results in pixels detecting temporal contrast ≥12.5% between frames, while n=4 captures motion events with temporal contrast ≥6.25%. The lowest contrast sensitivity tested was n=5, corresponding to a contrast sensitivity ≥3.1%, thus capturing all the spinning disk section transitions down to the lowest at 4% contrast. This is an improvement in contrast sensitivity of over a factor of 2 over the sensors reported in [2] [3] [4]. Reducing *n,* results in more low contrast transitions being missed, consistent with sensor sensitivity to only higher contrasts. With sensitivity to lower contrasts, smaller deviations in frame intensity including photon shot noise are captured as vision events. This can be problematic in the vision-triggered ToF modality, as noise in the vision frames will trigger a ToF acquisition continuously.

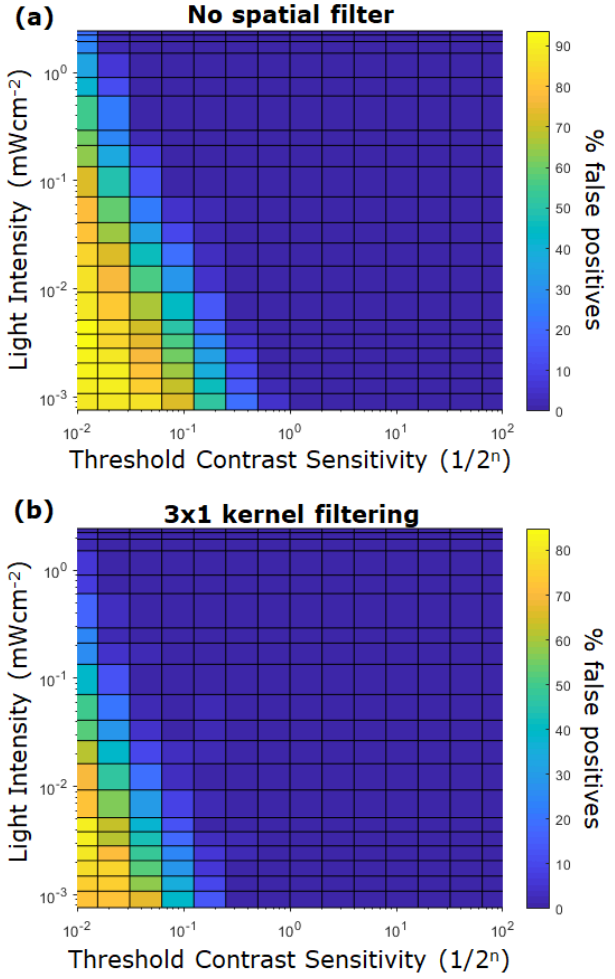

Fig. 18: (a) Percentage false positives with varying ambient background intensity and threshold with no spatial filtering. (b) Percentage false positives with varying ambient background intensity and threshold with 3x1 spatial filtering.

To filter out false positives and therefore prevent noise from triggering ToF acquisitions, the FPGA performs real-time spatial filtering to output events correlated across multiple pixels. The FPGA allows programmability of the spatial kernel pattern. A 2x1 and 3x1 kernel filtering were tested, where a vision motion event is recorded only if neighbouring pixels on the same row also output a motion event in the same frame. Linear filtering kernels were implemented as opposed to arrays, as the former favour real-time spatial filtering due to the rolling shutter readout of the sensor rows and optimise FPGA memory resources. A 3-bladed rotating fan was captured under 500 lux ambient illumination as shown in Fig. 17. The data is compressed into a single bit vision output by logically ORing the two vision bits to reduce the complexity of the spatial filtering implementation on the FPGA. By increasing the kernel size from a 2x1 to a 3x1 kernel, all noise-triggered vision events are successfully filtered out. In this implementation, kernel filtering was chosen to be implemented off-chip for ease of reconfigurability. As recently shown by [10], digital kernel-based motion spatial filtering can be integrated on-chip with negligible overhead in power consumption (<1 mW) and circuit area. The FPGA spatial filtering scheme used is composed primarily of low-power asynchronous digital circuits. It is therefore compatible with a possible future on-chip implementation in a motion-triggered 3D imaging sensor without affecting the system power budget compared to the current implementation.

The number of false-positive noise events in a scene varies with illumination and contrast sensitivity. Measuring the dependence of false-positives on the scene brightness and the contrast sensitivity is crucial to evaluate the probability that noise events will trigger ToF acquisitions when no motion has occurred. The plot in Fig. 18a shows the percentage of false positives detected in a frame capturing a scene with no motion. The sensor was exposed to a white static background. The background ambient light and threshold were gradually varied, recording the number of vision motion events in each acquired frame. With increased threshold sensitivity and lower background light, more false positives are recorded due to increasing shot-noise on the relative brightness changes. By implementing the 3x1 spatial filtering Fig. 18b shows an overall reduction in the percentage of false positives, improving the worst-case intensity and threshold contrast sensitivity conditions by almost 10% decrease in false positives, and removing false-events entirely in scenarios with contrast sensitivities above 12.5% at very low light intensities. The improvement in immunity to false positives allows the uninterrupted monitoring of a scene in MD-ToF until a vision event triggered by motion prompts the sensor to switch to measuring the scene depth. According to the dependence of false positives to the contrast sensitivity, the value of *n* is tuned in software to minimise the occurrence of false positives in response to the scene background light intensity. In MD-ToF

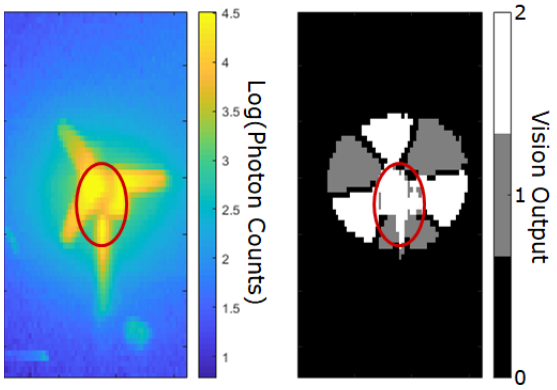

Fig. 19: Intensity and vision frames showing effect of saturation false-positives from pixel counter saturation (red circled region) due to high light intensity. The fan center and support post constantly output motion false events despite being static in motion or contrast. Saturation-triggered false-positives are distinguishable from true events as they do not change pixel location between consecutive frames.

mode the background light intensity is measured from the ToF histograms read out, using the bins not containing the return laser pulse in each pixel to determine the dynamic range of background light over the whole frame. The value of n is then adjusted and input to the sensor in response to the scene illumination dynamic range, according to Fig. 18b, to minimise spurious vision events while maximising contrast sensitivity.

Analogue implementations of vision sensors such as [2] [4] are affected by contrast variation due to the fixed pattern noise (FPN) caused by pixel-to-pixel mismatch in the storage node, amplifier, differencing and comparator circuits. This leads to threshold variations from one pixel to the other >2-4% which results in different contrast sensitivity at different pixels. Due to the digital nature of SPAD pixels, the FPN is limited in this sensor to mismatch in the passive quench transistor and variations in its bias voltage, setting the quench and recharge time of the SPAD, and variations in the high voltage applied at the SPAD cathode. Mismatches in these will result in SPADs having count rate variations from one pixel to another given the same incident illumination. No amplification is required due to the high gain of the SPAD and the frame differencing and comparison are entirely performed in the digital domain by the column-parallel processor. The FPN of the sensor was measured by diffusing a white LED subjecting the sensor FOV to spatially uniform illumination. The FPN was measured from the standard deviation of the counts to be 1.39%, smallest among vision sensors due to the low mismatch brought from a SPAD digital frame differencing design.

The dynamic range of the sensor evaluates its ability to detect motion events in a wide range of background illumination conditions. Intra-scene dynamic range is the ratio of the highest to lowest brightness within the same scene acquisition at which a high contrast motion event can still be detected [2]. The higher limit on the intra-scene sensor dynamic range is set by the bit-depth of the in-pixel counters. If either counter saturates, both counters will be locked by the overflow protection bit (counter MSB). In ToF modality, this feature ensures that the difference between counters is preserved when either counter saturates in bright conditions. In vision, this leads to the counter locking at its maximum and the other counter locking at 0 counts for the following exposure. This condition results in saturation-triggered vision false-positives as shown in Fig. 19. These false-positives are spatially and temporally correlated (high light level will saturate multiple adjacent pixels for multiple exposures) and can therefore be discriminated from noise-triggered false-positives. Saturating pixel false-positives can be used as a method to detect counter saturation and dynamically adjust the exposure to acquire the scene within the counter dynamic range. This is currently performed in software although could be ported to an FPGA or on-chip implementation.

In order to avoid saturation-triggered false positives the average count of the brightest signal C_{high} should not reach the counter full well C_{max}:

$$C_{high} + 3\sqrt{C_{high}} < C_{max} \tag{9}$$

where $3\sqrt{C_{high}}$ accounts for 99.7% (3σ) of the shot noise adding to C_{high}. With $C_{max}=2^{15}$, the maximum average count in the counters to ensure <0.15% probability of saturating the counters is $C_{high}=32229$. This is the upper bound of the intra-scene dynamic range.

In low-intensity background scene conditions, the shot noise variations results in many noise false-positives as shown by Fig. 18. To ensure a low probability of shot noise-triggered false-positives, the condition for the average pixel count C_{low} is:

$$\frac{3\sqrt{C_{low}}}{C_{low}} < \frac{1}{2^n} \tag{10}$$

where n is the contrast sensitivity coefficient and C_{low} is

$$C_{low} = (B + DCR) \times exposure \tag{11}$$

where B is the background illumination photon count rate. This ensures that the shot noise of the SPAD dark count rate (DCR) and the dim scene count rate does not overtake the contrast threshold causing an event. For the lowest level of contrast sensitivity coefficient used n=1 (>50% contrast sensitivity), the minimum value of C_{low} satisfying Eq. 10 is $C_{low}=37$. This is the lower bound of the dynamic range which ensures a low probability (<0.15%) of shot noise-triggered false-positives. The median DCR of the pixels was measured at 132 counts/s at room temperature. Assuming a short exposure (>100 fps) the

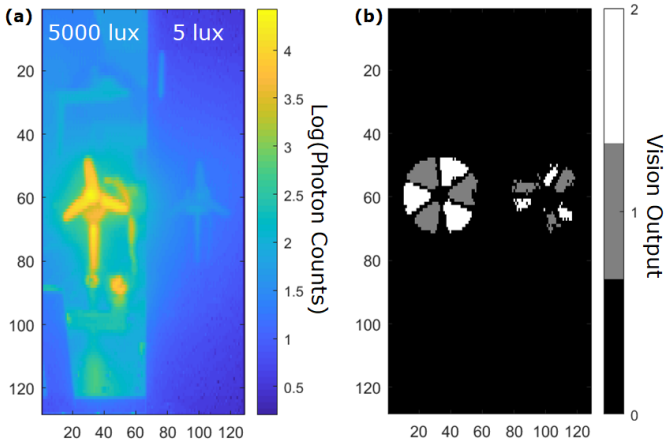

Fig. 20: (a) Intensity image (colorbar in log scale) of HDR scene (DCR subtracted) captured with the sensor in photon counting mode. (b) Vision frame at 60 dB scene dynamic range. 3x1 kernel spatial filtering removes false-positives.

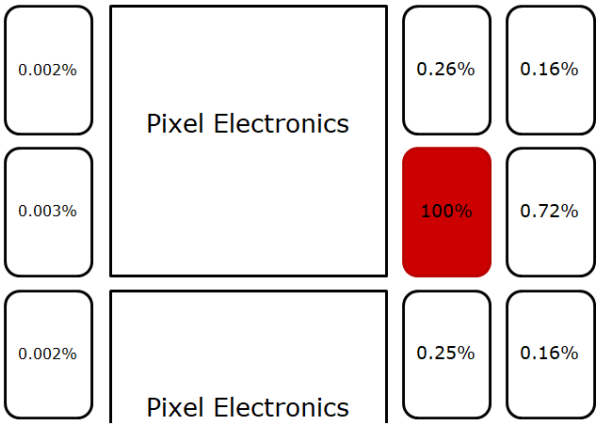

Fig. 21: Cross-talk probability of spatially adjacent SPADs from neighboring pixels. The SPAD highlighted in red is the reference SPAD relative to which cross-talk probability is measured.

lower bound of the dynamic range is therefore set by the minimum background illumination level B and the exposure such that $C_{low} \geq 37$.

Based on the values obtained for upper and lower bound, the intra-scene dynamic range is:

$$DR = 20 \log \frac{C_{high}}{C_{low}} = 58.8\ dB \qquad (12)$$

In practice, a higher dynamic range is possible up to around 69 dB as shown by the intensity range in Fig. 18 due to the 3x1 spatial kernel filter, removing shot-noise false-positives at lower C_{low}. The dynamic range reduces with higher values of contrast sensitivity n.

The intra-scene dynamic range was experimentally measured by subjecting the sensor to a scene split between high and low illumination. The method is the same used by [2]. Two halves of the scene were isolated by light-blocking panels with one side illuminated by a constant high power LED light. The brightness on the low illumination half of the scene was tuned via a current-controlled LED. The camera captured vision frames of both sides of the high dynamic range (HDR) scene in its FOV as shown in Fig. 20a. The bright side received 5 klux illumination and the dimmer side 5 lux. A 5 ms exposure was used. Noiseless acquisition of vision frames, while still resolving the shape of the fan blades on the darker side of the scene, were obtained as shown in Fig. 20b with a 60 dB dynamic range, closely matching the theoretical derivation.

Low noise motion detection is also achieved at lower brightness with longer exposures and higher brightness with shorter exposures, limited by the measured pixel dynamic range of 129 dB bounded by the maximum pixel count rate and the DCR of the 4 SPADs per pixel. Higher and lower scene intensity conditions however, require adjustment of exposure and uniformly-illuminated scenes (low intensity dynamic range) where the pixel counters will not saturate.

Low pixel cross-talk is important for both vision and ToF operation. High cross-talk can reduce the dynamic range by increasing the noise floor as well as causing spatial distortions in the image. Since SPADs are in close proximity by sharing the same well, the cross-talk between SPADs was measured by evaluating whether the avalanche of a SPAD is causing optical or electrical crosstalk triggering SPADs from neighboring

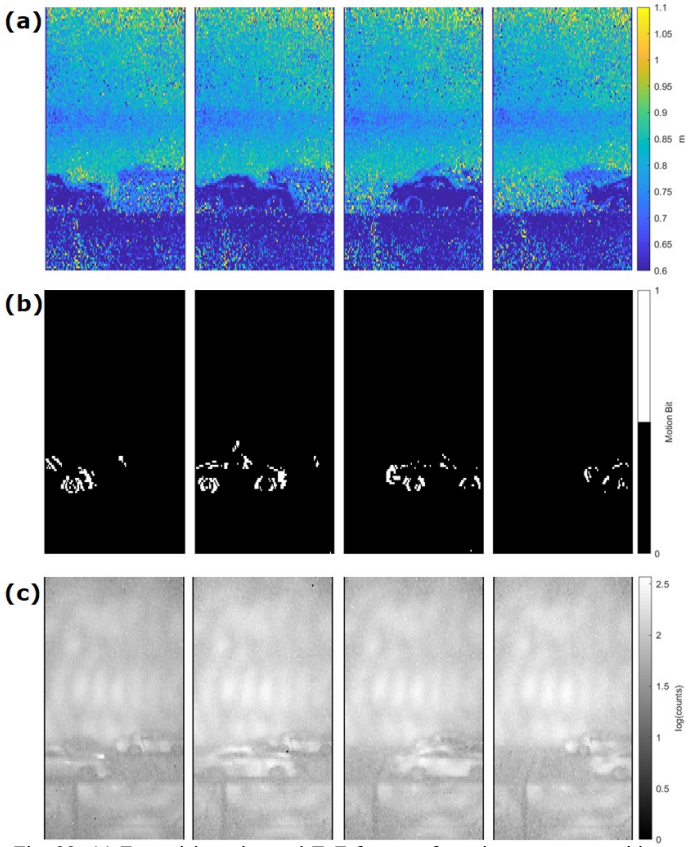

Fig. 22: (a) Four vision-triggered ToF frames of moving toy car transiting past a white background with red car in background. (b) Corresponding vision frames that triggered the ToF frames (vision output compressed to 1 bit). (c) Intensity frames extracted by summing 6 histogram bins in ToF

pixels. In this sensor SPADs cannot be individually disabled (only individually masked from the pixels counters), so direct measurement of cross-talk cannot be performed. However, an indirect cross-talk measurement as used in [23] was applied. A DCR map of every SPAD was created by capturing photon counting frames in the dark, successively masking the counts from 3 out of 4 SPADs from each pixel to obtain the DCR of individual SPADs. The top 1% highest DCR SPADs were located and the average rise in counts of the neighbouring pixels was measured. This gives an estimate of the cross-talk probability of one SPAD to adjacent pixel neighbours. The result of the cross-talk measurement is shown in Fig. 21. The highest average cross-talk probability of 0.72% occurs as expected between horizontally adjacent SPADs due to sharing a longer edge compared to diagonally or vertically adjacent SPADs or SPADs separated by pixel electronics. The peak probability measured is 1.78%, lower than measured in [23]. The pixel cross-talk probability is therefore low enough to consider any effects on noise and distortion negligible.

C. Motion-Triggered 3D Imaging

To demonstrate the motion-triggered 3D imaging function of the sensor a 100 fps video of the previous moving toy car was acquired in MD-ToF mode. Several frames of the video are shown in Fig. 22. The depth-maps are captured in response to the motion of the car following a vision frame. Two cars can be observed in the ToF frames: a moving white car in the foreground and a static red car in the background. ToF frame acquisitions are triggered by vision motion frames of the car in

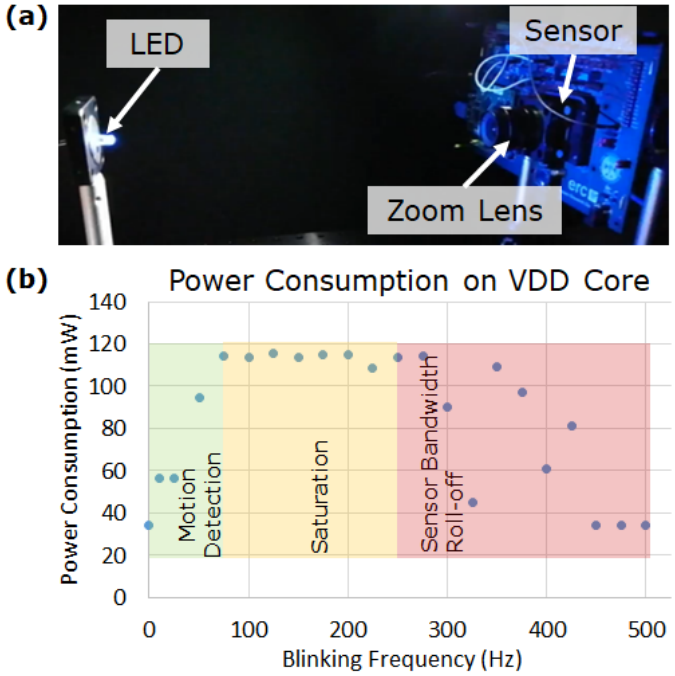

Fig. 23: (a) Experimental setup of frequency-modulated LED in front of sensor FOV operated in MD-ToF mode. (b) Power consumption of VDD core power supply with different LED flashing modulation frequencies.

the foreground only. Shadows cast by the white car over the red car also create motion events.

TABLE I. POWER CONSUMPTION BREAKDOWN OF THE SENSOR.

Power Supply	Vision Mode (steady-state)	Histogram Mode (steady-state)
VDD Core (1.1V)	33 mW	156 mW
IO Supply (3.3V)	1.2 mW	8.3 mW
SPAD VHV (16V) @ 100 klux	21 mW	21 mW
Total	55 mW	185 mW

Table I shows the power consumption breakdown of the sensor for the different power supplies in the vision and histogram modes at 100 klux background illumination and running the sensor at 500 fps. The power consumption is measured to be 70% less in vision motion detection modality than for ToF histogram acquisition. This results in a motion activity-dependent power consumption in MD-ToF mode.

An experiment was run to observe the variation in power consumption of the sensor in MD-ToF mode according to scene motion activity. The sensor was operated at 500 fps in MD-ToF mode. A frequency-modulated flashing LED was placed in front of the sensor to simulate a motion event through brightness changes. The power of the VDD core supply was measured at different LED modulation frequencies as shown in Fig. 23. The graph shows a power consumption variation consistent with the steady-state consumption at different switching duty cycles between vision operation and histogram acquisition. Three regions of operation are highlighted. The first, occurs between 0 and 75 Hz, where the sensor linearly increases the number of histogram acquisitions due to more frequent motion events. Between 100 Hz and 250 Hz the sensor detects a motion in every vision frame and therefore toggles consecutively between the two modes, which results in saturation of the power consumption. After 250 Hz the LED flashing exceeds the vision bandwidth and the brightness changes are averaged over the sequential vision frames. The

TABLE II. STATE OF THE ART TOF AND VISION IMAGERS COMPARISON

Parameter	ToF Sensors				
	This Work	**[14]**	**[16]**	**[12]**	**[13]**
Technology (nm)	40	65 BSI	130	40/90 BSI	180
Detector	SPAD	QEM-PD	SPAD	SPAD	SPAD
Resolution	128x128	1024x 1024	128x96	64x64	252x144
Pixel Pitch (µm)	40 x 20	3.5	44.65	38.4	28.5
Fill Factor (%)	13	~100	3.17	51	28
Frame Rate (fps)	500	30	20	760	30
Motion Detection	Yes	No	No	No	No
Power Cons. (mW)	185	650	40	77.6	2540
Range (m)[1*]	1.50 (min) 48 (max) 3 (meas.)	0.4-4.8	0-45	0-50	2-50
ToF Accuracy (% range)	0.6	0.2	6	0.34	0.17
Norm. Power Cons.[2*]	6.7-22.5	20.6	163	24.9	2333
Norm. Idle Sensor Data Rate[3*] (KB/s)	120	400	400	8960	3600
Parameter	**Vision Sensors**				
	This Work	**[9]**	**[8]**	**[3]**	**[2]**
Vision Type	Frame Comp.	Frame Comp.	Frame Comp.	DVS-AER	DVS-AER
Techn. (nm)	40	65	90/40 BSI	90	350
Detector	SPAD	CIS	CIS	CIS	CIS
Resolution	128x128	32x20	160x154	640x480	128x128
Pixel Pitch (µm)	40 x 20	1.5	1.5	9	40
Fill Factor (%)	13	-	-	-	8.1
Frame Rate (fps)	500	170	10	-	-
Dynamic Range (dB)	69	64.3	96	80	120
Min. Contrast (%)[4*]	4	-	-	9	13
FPN (%)	1.39	-	-	-	2.1
3D Imaging	Yes	No	No	No	No
Power Cons. (mW)[5*]	55-185	4.5	1.1	27-50	30

[1*]Minimum range is for 2 ns time gates and maximum range is for 64 ns time gates with 5 bins of non-ambiguous range. Measured 3 m range using 4 ns time gates.
[2*]Normalized power consumption = Power/Frame Rate/Number of Pixels.
[3*]Normalized to 128x128 resolution and 20fps frame rate for all works.
[4*]Minimum measured contrast in vision modality. 12.5% contrast limit used for vision-ToF experiments to avoid noise triggering ToF capture after every vision frame. Processor minimum contrast hardware limit is 0.8%.
[5*]Power consumption measured at 100 klux illumination and 500 fps.

results of this experiment show the efficiency of the sensor in switching to a high power modality when detecting a motion event and remaining in a low power and low data rate operation in idle activity. This modality would be beneficial in industrial automation applications, for example in classification, tracking and sorting of objects transitioning on a conveyor belt. Surveillance and assisted living 3D cameras tracking movement and location of people would use their energy more efficiently by limiting 3D image capture triggered only by motion events such as detecting falls or entrance in forbidden areas.

IV. Comparison to State of the Art

Table II compares the sensor with state of the art vision and ToF sensors. The sensor is distinguished in its motion-triggered ToF functionality, which results in a low activity-dependent power consumption when compared to other sensors for similar resolution and frame rate. The sensor data rate in idle motion detection state is in some cases orders of magnitude smaller than high data rate ToF sensors, which leads to a reduced power requirement on off-chip data processing. The sensor is also the first implementation of a SPAD vision sensor. The entirely digital design results in improved sensitivity to less than half the reported minimum contrast [2] [3] and lower contrast threshold FPN than the state of the art vision sensors [2] [4]. Background subtraction vision cameras like [6] [7] [10] are the lowest power consuming vision technique; however, background subtraction algorithms require per pixel threshold adjustment and tuning of coefficients to set the learning rate of the algorithms to respond to intensity changes. The more complex low-pass filtering [7] and recursive approaches [24] [25] are not as easily portable from the analogue to the digital domain without increasing the pixel pitch or additional per-pixel memory resources. The sensor dynamic range is comparable to most frame comparison-based vision sensor, with the exception of sensors using spatial aggregation to increase their dynamic range at the cost of lower resolution [8]. Higher dynamic range techniques taking advantage of logarithmic acquisition in DVS are incompatible with SPAD digital photon counting. However, the SPAD dynamic range is comparable to DVS logarithmic sensors and could be fully exploited with the integration of on-chip auto-exposure feedback or in-pixel high dynamic range counters [12].

V. Conclusion

Embedded on-chip processing integrated in advanced deep submicron technologies allows the integration of multiple imaging modes on the same sensor. It has been demonstrated how vision and ranging functionalities can share resources at the pixel level to preserve imaging resolution. By interleaving vision and ToF acquisitions the presented camera achieves motion-triggered depth imaging, thus opening opportunities for 3D imaging applications in low-power IoT systems.

Acknowledgment

The authors thank STMicroelectronics for the fabrication of the integrated circuit.

References

[1] G. Gallego, T. Delbruck, G. Orchard, C. Bartolozzi, B. Taba, A. Censi, S. Leutenegger, A. Davison, J. Conradt, K. Daniilidis, and D. Scaramuzza, "Event-based vision: A survey, " *arXiv preprint arXiv:1904.08405*, 2019.

[2] P. Lichtsteiner, C. Posch and T. Delbruck, "A 128 x 128 120db 30mW asynchronous vision sensor that responds to relative intensity change," *2006 IEEE International Solid State Circuits Conference - Digest of Technical Papers*, San Francisco, CA, 2006, pp. 2060-2069.

[3] B. Son *et al*., "4.1 A 640×480 dynamic vision sensor with a 9µm pixel and 300Meps address-event representation," *2017 IEEE International Solid-State Circuits Conference (ISSCC)*, San Francisco, CA, 2017, pp. 66-67.

[4] C. Brandli, R. Berner, M. Yang, S. Liu and T. Delbruck, "A 240 × 180 130 dB 3 µs Latency Global Shutter Spatiotemporal Vision Sensor," in *IEEE Journal of Solid-State Circuits*, vol. 49, no. 10, pp. 2333-2341, Oct. 2014.

[5] N. Massari, M. De Nicola and M. Gottardi, "A 30µW 100dB contrast vision sensor with sync-async readout and data compression," *2010 Proceedings of ESSCIRC*, Seville, 2010, pp. 138-141.

[6] M. Gottardi, N. Massari and S. A. Jawed, "A 100 uW 128x64 Pixels Contrast-Based Asynchronous Binary Vision Sensor for Sensor Networks Applications," in *IEEE Journal of Solid-State Circuits*, vol. 44, no. 5, pp. 1582-1592, May 2009.

[7] N. Cottini, M. Gottardi, N. Massari, R. Passerone and Z. Smilansky, "A 33uW 64x64 Pixel Vision Sensor Embedding Robust Dynamic Background Subtraction for Event Detection and Scene Interpretation," in *IEEE Journal of Solid-State Circuits*, vol. 48, no. 3, pp. 850-863, March 2013.

[8] O. Kumagai, A. Niwa, K. Hanzawa, H. Kato, S. Futami, T. Ohyama, T. Imoto, M. Nakamizo, H. Murakami, T. Nishino, A. Bostamam, T. Iinuma, N. Kuzuya, K. Hatsukawa, F. Brady, W. Bidermann, T. Wakano, T. Nagano, H. Wakabayashi, Y. Nitta, "A 1/4-inch 3.9Mpixel low-power event-driven back-illuminated stacked CMOS image sensor," *2018 IEEE International Solid - State Circuits Conference - (ISSCC)*, San Francisco, CA, 2018, pp. 86-88.

[9] K. D. Choo, L. Xu, Y. Kim, J.-H. Seol, X. Wu, D. Sylvester, D. Blaauw, "5.2 Energy-efficient low-noise CMOS image sensor with capacitor array-assisted charge-injection SAR ADC for motion-triggered low-power IoT applications," *2019 IEEE International Solid- State Circuits Conference - (ISSCC)*, San Francisco, CA, USA, 2019, pp. 96-98.

[10] Y. Zou, M. Gottardi, M. Lecca and M. Perenzoni, "A Low-Power VGA Vision Sensor with Event Detection through Motion Computation based on Pixel-Wise Double-Threshold Background Subtraction and Local Binary Pattern Coding," *ESSCIRC 2019 - IEEE 45th European Solid State Circuits Conference (ESSCIRC)*, Cracow, Poland, 2019, pp. 97-100.

[11] S. Yoshimura, T. Sugiyama, K. Yonemoto and K. Ueda, "A 48 kframe/s CMOS image sensor for real-time 3-D sensing and motion detection," *2001 IEEE International Solid-State Circuits Conference. Digest of Technical Papers. ISSCC (Cat. No.01CH37177)*, San Francisco, CA, USA, 2001, pp. 94-95.

[12] S. W. Hutchings, N. Johnston, I. Gyongy, T. Al Abbas, N. A. W. Dutton, M. Tyler, S. Chan, J. Leach, R. K. Henderson, "A Reconfigurable 3-D-Stacked SPAD Imager With In-Pixel Histogramming for Flash LIDAR or High-Speed Time-of-Flight Imaging," in *IEEE Journal of Solid-State Circuits*, vol. 54, no. 11, pp. 2947-2956, Nov. 2019.

[13] C. Zhang, S. Lindner, I. M. Antolović, J. M. Pavia, M. Wolf and E. Charbon, "A 30-frames/s, 252 x 144 SPAD flash LiDAR with 1728 dual-clock 48.8-ps TDCs, and pixel-wise integrated histogramming," in *IEEE Journal of Solid-State Circuits*, vol. 54, no. 4, pp. 1137-1151, April 2019.

[14] C. S. Bamji, S. Mehta, B. Thompson, T. Elkhatib, S. Wurster, O. Akkaya, A. Payne, J. Godbaz, M. Fenton, V. Rajasekaran, L. Prather, S. Nagaraja, V. Mogallapu, D. Snow, R. McCauley, M. Mukadam, I. Agi, S. McCarthy, Z. Xu, T. Perry, W. Qian, V.-H. Chan, P. Adepu, G. Ali, M. Ahmed, A. Mukherjee, S. Nayak, D. Gampell, S. Acharya, L. Kordus, "1Mpixel 65nm BSI 320MHz demodulated TOF Image sensor with 3µm global shutter pixels and analog binning," *2018 IEEE International Solid - State Circuits Conference - (ISSCC)*, San Francisco, CA, 2018, pp. 94-96.

[15] C. Niclass, M. Soga, H. Matsubara, M. Ogawa and M. Kagami, "A 0.18-µm CMOS SoC for a 100-m-Range 10-Frame/s 200 × 96-Pixel Time-of-Flight Depth Sensor," in *IEEE Journal of Solid-State Circuits*, vol. 49, no. 1, pp. 315-330, Jan. 2014.

[16] R. J. Walker, J. A. Richardson and R. K. Henderson, "A 128×96 pixel event-driven phase-domain ΔΣ-based fully digital 3D camera in 0.13µm CMOS imaging technology," *2011 IEEE International Solid-State Circuits Conference*, San Francisco, CA, 2011, pp. 410-412.

[17] J. Noraky and V. Sze, "Low power depth estimation for time-of-flight imaging," *2017 IEEE International Conference on Image Processing (ICIP)*, Beijing, 2017, pp. 2114-2118.

[18] F. M. Della Rocca, H. Mai, Sam. W. Hutchings, T. Al Abbas, A. Tsiamis, P. Lomax, I. Gyongy, N. A. W. Dutton, R. K. Henderson, "A 128 × 128 SPAD Dynamic Vision-Triggered Time of Flight Imager," *ESSCIRC 2019 - IEEE 45th European Solid State Circuits Conference (ESSCIRC)*, Cracow, Poland, 2019, pp. 93-96.

[19] S. Pellegrini, B. Rae, A. Pingault, D. Golanski, S. Jouan, C. Lapeyre, B. Mamdy, "Industrialised SPAD in 40 nm technology," *2017 IEEE International Electron Devices Meeting (IEDM)*, San Francisco, CA, 2017, pp. 16.5.1-16.5.4.

[20] F. Mattioli Della Rocca, T. A. Abbas, N. A. W. Dutton and R. K. Henderson, "A high dynamic range SPAD pixel for time of flight imaging," *2017 IEEE SENSORS*, Glasgow, 2017, pp. 1-3.

[21] L. Frey, L. Masarotto, P. Gros D'Aillon, C. Pellé, M. Armand, M. Marty, C. Jamin-Mornet, S. Lhostis, and O. Le Briz, "On-chip copper–dielectric interference filters for manufacturing of ambient light and proximity CMOS sensors," Appl. Opt. 53, 4493-4502 (2014).

[22] I. Gyongy, N. Dutton, and R. K. Henderson, "Single-Photon Tracking for High-Speed Vision," *Sensors,* Basel, Switzerland, 2018, 18(2), 323.

[23] I. Gyongy *et al*., "A 256 x 256, 100-kfps, 61% Fill-Factor SPAD Image Sensor for Time-Resolved Microscopy Applications," in *IEEE Transactions on Electron Devices*, vol. 65, no. 2, pp. 547-554, Feb. 2018.

[24] A. Manzanera and J. Richefeu, "A Robust and Computationally Efficient Motion Detection Algorithm Based on Sigma-Delta Background Estimation," in *ICVGIP 2004, Proceedings of the Fourth Indian Conference on Computer Vision, Graphics & Image Processing,* 46-51, December 2004.

[25] A. Verdant, A. Dupret, H. Mathias and P. Villard, "Adaptive thresholding for motion detection in a CMOS image sensor," *2007 Conference Record of the Forty-First Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, CA, 2007, pp. 495-499.